Journal of Internet Services and Applications, 2024, 15:1, doi: 10.5753/jisa.2024.3980

Data quantitative and qualitative study in Brazilian Open Data Portals

Shirlei L. O. do Carmo [**Universidade Federal do Rio Grande do Sul** | ***slocarmo@inf.ufrgs.br***]
Claudio F. R. Geyer [**Universidade Federal do Rio Grande do Sul** | ***geyer@inf.ufrgs.br***]
Julio C. S. dos Anjos [**Federal University of Ceara Campi Itapaje | PPGETI** | ***jcsanjos@ufc.br***]

Institute of Informatics, Universidade Federal do Rio Grande do Sul, Av. Bento Gonçalves, 9500, 91501-970 Porto Alegre, RS Brazil. Full list of authors' information is available at the end of the article.

Received: 28 December 2023 • **Accepted:** 21 April 2024 • **Published:** 25 June 2024

Abstract Open data is a concept attributed to sharing data with anyone, and in addition to being accessed, this data can be manipulated and redistributed. The optimized and interchangeable use of open data can lead to so-called open innovation, which can be understood as the crossing of information between different organizations, to generate more complete and innovative systems and solutions. Despite the clear benefit for society, there are major challenges highlighted in different studies for its implementation, such as the lack of promotion of open data, the lack of standardization in data availability, as well as the lack of complete and updated information, among others. This study uses an available reproducible methodology, to show, through different dimensions, the open data panorama in Brazil, which indicates that there are many opportunities for improvement, in categories such as standardization of data exposure and its licenses, update rate, and, due to the absence of some data, the lack of promotion of open data.

Keywords: Open Data; Transparency; Open Knowledge; Ckan; Open Innovation

1 Introduction

The knowledge generated through data analysis is a necessity and a challenge in a scenario where different computerized systems often generate it unbridled and unplanned. It is a necessity because the benefits to society are indisputable, such as knowledge and oversight of public spending McDermott [2010], access to health and environment data by population, planning of public civil works, forest fires in the Amazon, use in research and development of applications for the benefit of society Chen and Jakubowicz [2015] Guo *et al*. [2019] Zhang and Yue [2016], among others.

The challenges are related to the rare promotion and awareness of these benefits so that there is the effective use of data Janssen *et al*. [2012]. As for availability and sharing, there is a lack of awareness of the importance of standardization in opening data so that it is reusable and interchangeable in different opportunities. Other difficulties perceived by Machado *et al*. [2019] regarding open data in education were verified during the study performed in this paper, which are dispersion, unclear licensing, insufficient standardization of data, and lack of incentives and infrastructure for data sharing.

These challenges are in accordance with the main ideas promoted by the Open Knowledge Foundation (OKFN), a non-profit, non-partisan civil society organization, when defining the concept of "open", which are: knowledge as a common good, being possible for anyone to use and participate in its construction; and, whether computerized or not, the systems must be "interoperable", which means maximizing their capacity to communicate transparently and to connect with other systems Foundation [2023].

Despite the great challenges, open data has great potential, and different initiatives seek to promote it, such as the OKFN, which supports and guides the use and sharing of data. Through OKFN, standardization initiatives in the availability of data are encouraged, and, currently, on its page, it is possible to verify CKAN (a data management system) as a suitable tool for sharing, making available, and searching for open data. On the CKAN website, it is possible to verify that countries such as the United States, Canada, Germany, and Australia already use this tool in their open data portals.

The benefits go beyond knowledge of open data by the population and government transparency. Forbes, a renowned business and economics magazine, reported in 2020 that using open data is essential for projects that seek to create smart cities Arbex [2020]. In short, smart cities can be understood as technological solutions — applied in cities, neighboring regions, and rural areas — in order to build sustainable open environments and user-driven innovation ecosystems Domingue *et al*. [2011]. According to the magazine, London, the city that created the first open data store in the world — the London Datastore[1] — and a pioneer in the adoption of smart cities, had at the time the objective of sophisticating its public data sharing in the following years, intending to solve problems generated by the population growth in urban centers. According to the United Nations[2], the world population will be 68% urban by 2050 DESA [2018], so the relevance of open data for technological initiatives such as smart cities has proven to be fundamental, as seen in Adje *et al*. [2023] work.

[1] https://data.london.gov.uk/dataset
[2] https://www.un.org/en/

In Brazil, the incentive to public access information is encouraged by the Federal Government through laws such as the *Access to Information Law* (LAI), thus, standardization policies for data opening Federal [2023b], and construction of National Data (INDA) Federal [2023a] are implemented. All of these initiatives have the ultimate objective of guiding the dissemination and sharing of data and public information in an open format. However, despite the encouragement through federal laws and initiatives[3] to monitor the opening of data in the states of Brazil, there are few studies on the quality and status of the data available on the portals, and comprehensive long-term initiatives that make it possible to monitor the evolution of the opening of this data are still scarce. This creates uncertainty about the feasibility of using open data in Brazil since the quality of the data is not known.

Due to the scenario shown, the following contributions of this work will be highlighted:

- A current picture of the health of open data in Brazil, demonstrating through qualitative analysis how dimensions/categories relevant to society — such as *'Public Spending'* — are available, updated, and formatted according to standards used in open data portals by the world. Also, to support the panorama demonstration of open data in Brazil, a quantitative analysis was carried out on the open data portals of the states in Brazil. With this, it was verified which states have open data portals, what amounts of data were found, what variety of subjects/themes on the portals, and, finally, we sought to correlate variables such as 'quantity of datasets' with social variables such as per-capita income, and population quantity.
- The study will provide government institutions and Brazilian states with evidence of the gaps existing in the distribution of open data made available, enabling improvements in the quality of exposure;
- Regarding the academic community, according to Pareja-Lora *et al*. [2019], the publication of scientific data under open resources - in this case, Open Data - has become routine in modern research and the Open Data movement in linguistics - as well as in all areas of study of science, computing, and humanities – is based on three primary motivations: responsibility, reproducibility, and reuse. For this purpose, the framework provided can be used to evaluate these motivations, checking through the indicators presented in the study, such as: Does the data exist? Does the data from this research exist? Publicly available? Is the data provided on a timely and up-to-date basis? So, gaps exist, and insights for improvement can be mapped and applied. For instance, *Is the data machine-readable?* checks whether a machine can read the dataset in question and, therefore, is capable of reuse or use in automation. The ideal data format can easily be substituted for each particularity in the available framework.

The article was organized as follows. Section 2 contains the related works, with the latest available analyses made in Brazil and institutions responsible for the reports with its methodologies. The section 3 explains the flow of data analysis, from ingestion to exploration and construction of results. Section 4 shows the technologies used, the way of exposing the data, the datasets to be evaluated, the CKAN API calls used, and the metrics evaluated. The results found in each dimension will be demonstrated in Section 5, in addition to the difficulties in exploring the data in each one. Finally, Section 6 summarizes the main conclusions of this work.

2 Background and Related Works

The Open Knowledge Foundation created the Open Data Index (ODI) initiative, a pioneering initiative in promoting transparency, helping to evaluate policies, identify bottlenecks, and guide municipalities to improve their open data policies Index [2018]. The Index evaluates not only federal but also municipal governments, acting to ensure the necessary scalability. ODI was brought to Brazil through a partnership between the Directorate of Public Policy Analysis (DAPP) of the Getúlio Vargas Foundation (FGV) - FGV is a Brazilian institution known worldwide as a reference in teaching and research - and Open Knowledge Brasil (OKBR) [4]. The ODIs have been launched in Brazil since 2017, with 2018 being the last year of updates, according to the FGV website.

In the **Figure 1**, there is an overview of the 2018 Open Data Index Score. Which, according to the publication, evaluates the adequacy of the data made available by the government to the transparency criteria used in several countries around the world and through the "% open". The percentage of the evaluated datasets that meet all the criteria of the methodology is calculated.

According to the study, at the time, São Paulo was the city, among the eight evaluated, that presented the highest score and also the highest %open. This means that the city was the most successful both in disclosing 100% open bases and in bringing its public bases closer to the Open Definition criteria. However, the challenges remain the same today (as will be seen after this work), as the study adds that despite the good results in São Paulo, it does not mean that all challenges have already been met since less than half of the bases of data evaluated met the full criteria of the ODI.

Cities	Score	Cities	% Open
São Paulo	84%	São Paulo	47%
Rio de Janeiro	75%	Belo Horizonte	35%
Belo Horizonte	73%	Rio de Janeiro	29%
Porto Alegre	68%	Brasília	29%
Brasília	68%	Porto Alegre	23%
Salvador	55%	Uberlândia	17%
Uberlândia	53%	Natal	11%
Natal	43%	Salvador	5%

Figure 1. ODI city scores according to score and %Open. Source: FGV/DAPP e Open Knowledge Brasil, 2018

The Open Definition project Foundation [2023] affirms

[3] https://centralpaineis.cgu.gov.br/visualizar/dadosabertos

[4] https://ok.org.br/

that the term "open" in the context of "open data" and "open content" means that data can be freely accessed, used, modified, and shared by anyone, with anyone. Subject, at most, to requirements that preserve provenance and openness.

Open Knowledge Brazil also defines the main conditions for data to be considered open. In short, (1) availability and access, that is, It must be fully available and only at the cost of copying; (2) also in a convenient and changeable format; (3) its reuse and redistribution must be possible, that is, in addition to being possible to reuse, it must be possible to combine it with another set of data; (4) and universal participation, that is, everyone must be able to use it, without any discrimination against people, groups or relative fields of action (such as only for non-profit or educational purposes).

3 Data Analysis Model

The article's data analysis model has the following steps (1) a manual search for Brazilian states or institutions that have an open data portal, (2) checking whether the data is exposed through the CKAN data management system - essential, because even if the data portal has a data exposure standard if it does not meet a higher exposure standard - such as CKAN - it will not be possible to apply automation along with other portals - (3) the URLs of the study portals are stored in text files, which will be used in all data analysis automation.

After collecting the URLs, the category to be evaluated is identified. For example, it will be verified if the 'data exists', if it is 'up-to-date', what the ' data license' is, etc. After organizing the category to be validated, the CKAN API method is performed. The result is collected and stored.

The high-level data flow can be seen detailed in **Figure 2**, which follows these steps:

(a) The URL of the open data portal is checked manually, for validation of the data exposure mode, CKAN usage is a requirement for analysis;
(b) The portal URL is manually inserted in a file, along with the CKAN API function. Then, the Python script must be started, initializing a function to search for resources in the portal. In the next step, the script will search for the URL of the portals to be analyzed in the file. The script concatenates the portal URL with the *CKAN API action*, finally, the URL for the HTTP request is assembled and requests are initiated according to the JSON contract CKAN [2023];
(c) After that, the script will create a list of resources in order to be stored in an array, and a second validation of the availability of the resource is performed. In this context, it is verified if the resource returned HTTP status 200, which is the HTTP response status that means success.

4 Methodology

This study was developed by joining different libraries to provide a technologically consistent query framework. During this evaluation, an exhaustive government dataset search

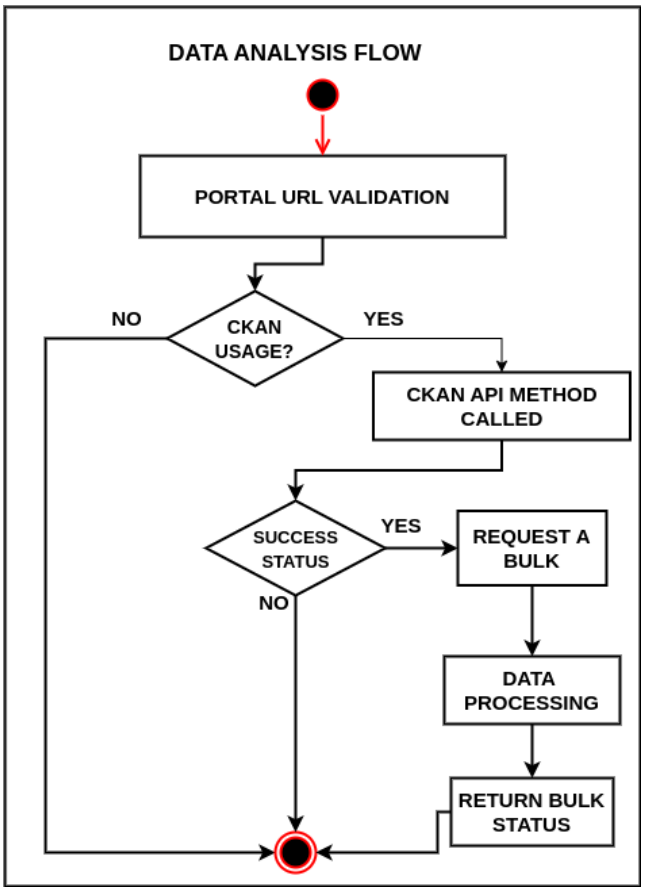

Figure 2. Data Analysis Model. Source: Author

from different Brazilian government ministries was accomplished.

4.1 Technologies

The technologies used were:

- The Anaconda distribution, which is an open-source platform for Data Science AnacondaOrg [2023], provides the prototype development environment;
- The web-based development environment was JupyterLab, which supports a wide range of workflows in data science, scientific computing, and machine learning JupyterOrg Community [2023]. In this work, it was chosen because it provides the possibility to run code in cells individually, which allows a gain of performance, for example,
- The programming language used was Python, as it is easy to integrate with other technologies, runs in different environments, and is simple to write PythonOrg [2023];
- The Pandas library for data analysis and manipulation developed for the Python language PandasOrg [2023];
- The Matplotlib library for plotting (image production) developed for the Python language MatplotlibOrg [2023].

The code used is available in the GitHub open source repository [5] and the guide to install and run the *Jupyter Notebook* can be found on *Jupyter's* web page JupyterOrg [2023].

[5] https://github.com/shluh/ufrgs-open-data-analysis

API Function	Description
/action/package_list	returns the set of datasets available by platform
/action/group_list	returns the groups that are contained in the datasets
/action/tag_list	returns the breakdown of tags by dataset
/action/package_show?	returns the representation of the dataset specified in param 'id'
/action/package_search?	to search for datasets (packages) matching the search query

Table 1. JSON-formatted lists of datasets

4.2 Datasets and Metrics

Open Source platforms have facilitated enormously the labor of institutions involved in Open Data initiatives. For instance, CKAN, based on Python technology, is the most widely used open-source platform to support Open Data portals Nogueras-Iso *et al*. [2021].

The open-source solution CKAN has grown to be the de facto standard in the public sector for creating open data catalogs but is increasingly adopted by private companies, too Kirstein and Bohlen [2022].

Thus, the technology used to extract the data was CKAN's Action API. It was also chosen because government agencies and open data portals in the states of Brazil, which were available for evaluation, mostly used the CKAN data management system at the time of the study.

Therefore, only datasets that expose data through this API - or derivations like DKAN - were analyzed. The technical framework used was *GET-able API functions*, with features by dataset, group, and resources, see **Table 1**.

According to the scope of this study, the official open data portals will be analyzed: Brazil and Federal District (see **Table 2**) - on the date the article was written, the Brazilian portal was undergoing a reformulation, and for compatibility with CKAN, the legacy URL of the portal was used *legado.dados.gov.br*, and not the main one *dados.gov.br*; each of the 26 states of Brazil (see **Table 3**).

Transparency portals were considered and analyzed individually - when the open data portal was not found - but disregarded when they did not use CKAN to expose the data. According to the Brazilian open data portal, the difference between the transparency portal and the open data portal is that the transparency portals have the objective of increasing the control of government expenses and revenues, and open data portals, on the other hand, have a different objective: to be the single point of reference for searching and accessing Brazilian public data on any and all subjects or categories. It is a simplified service that organizes and standardizes access to public data, focusing on data reuse and modern technologies.

In the absence of the state's open data portal and transparency portal, a search was made for the capital's open data portal.

Table 2. Brazilian and Federal District open data portal

District	Acronym	Data Portal URL
Distrito Federal	DF	http://www.dados.df.gov.br/
Brazil	BR	https://legado.dados.gov.br/

Table 3. Open data portal by state.

State	*Open Data Portal	CKAN	Population *(in millions)*
Acre	Yes	No	830,018
Alagoas	Yes	Yes	3,127,683
Amapá	Yes	No	733,759
Amazonas	Yes	No	3,941,613
Bahia	Yes	Yes	14,141,626
Ceará	Yes	No	8 794,957
Espírito Santo	Yes	Yes	3,833,712
Goiás	Yes	Yes	7,056,495
Maranhão	Yes	No	6,775,805
Mato Grosso	Yes	No	3,658,649
Mato Grosso do Sul	Yes	Yes	2,757,013
Minas Gerais	Yes	Yes	20,538,718
Pará	Yes	No	8,121,025
Paraíba	Yes	No	3,974,687
Paraná	Yes	No	11,444,380
Pernambuco	Yes	Yes	9,058,931
Piauí	Yes	No	3,271,199
Rio de Janeiro	Yes	No	16,054,524
Rio Grande do Norte	Yes	No	3,302,729
Rio Grande do Sul	Yes	Yes	10,882,965
Rondônia	Yes	No	1,581,196
Roraima	Yes	No	636,707
Santa Catarina	Yes	Yes	7,610,361
São Paulo	Yes	Yes	44,411,238
Sergipe	Yes	No	2,209,558
Tocantins	Yes	No	1,511,460

* or Transparency Portal

5 Results

For the qualitative analysis, the analyzed dimensions/categories were based on the Global Open Data Index (OKFN) — an annual effort to measure the state of open government data worldwide — methodology. In other words, dimensions/categories represent important data that is relevant to civil society at large. Therefore, the validation methodology was based on the same model. The dimensions are the following: National Statistics, Government Budget, Government Spending, Legislation, Election Results, National Map Pollutant Emissions Company Register Location datasets, Government procurement tenders (past and present); Water Quality, Weather forecast, Land Ownership; Transport Timetables; Health Performance. There will be no ranking and counting of points produced in OFKN, but the analysis result will be shown for all datasets available in open data portals.

The analysis covered the following questions, listed and described in accordance with OKFN[6]:

- Does the data exist? Does the data exist at all? The data can be in any form (paper or digital, offline or online, etc.);
- Is data in digital form? This question addresses whether

[6]https://opendatahandbook.org/

the data is in digital form (stored on computers or digital storage) or if it is only in e.g., paper form;
- Publicly available? This question addresses whether the data is ”public”. This does not require it to be freely available but does require that someone outside of the government can access it in some form. If a freedom of information request or similar is needed to access the data, it is not considered public;
- Is the data available for free? This question addresses whether the data is available for free or if there is a charge;
- Is the data available online? This question addresses whether the data is available online from an official source;
- Is the data machine-readable? The data is machine-readable if it is in a format easily structured by a computer. Data can be digital but not machine-readable. For example, consider a PDF document containing tables of data. These are digital but are not machine-readable because a computer would struggle to access the tabular information;
- Available in bulk? *In bulk* means, per *Open Definition*[7] that the data should be available as a complete set. And, if it has a register *i.e.* collected under the statute, the entire register should be available for download. So, the *data is available in bulk* if the whole dataset can be downloaded or accessed easily. On the other hand, it is considered non-bulk if the citizens are limited to just getting parts of the dataset (for example, if restricted to querying a web form and retrieving a few results at a time from a very large database);
- Openly licensed? This question addresses whether the dataset is open as per *http://opendefinition.org*. It needs to state the terms of use or license that allow anyone to freely use, reuse, or redistribute the data (subject at most to attribution or share-alike requirements). It is vital that a license is available (if there is no license, the data is not openly licensed). Open licenses which meet the requirements of the Open Definition are listed at *http://opendefinition.org/licenses/*;
- Is the data provided on a timely and up-to-date basis? This question addresses whether the data is up to date and timely - or long delayed.

For the quantitative analysis, the work of Barbosa *et al*. [2014] was used as a reference. Therefore, the country's states were validated from a statistical point of view. Then the following were reported: quantities of data sets, description of parameters and their meanings, and variety of information found in the open data portals of the states of Brazil during the experiments. This, in turn, makes it possible to monitor the evolution — over time — of the resources available on data portals, as was done in the work of Gharawi *et al*. [2019], who used the content analysis approach for this purpose. Therefore, the points of analysis are:

- Quantities of Groups/Theme per platform: aiming to find the variety of subjects available on data portals since groups in CKAN are used to create and manage collections of datasets. It can also help identify the level of organization in the open data portal, as groups are a simple way to help users find and search their own published datasets;
- Number of datasets per platform: aiming to show what gaps exist in availability; for example, the data portal ”X” does not provide any dataset with information on public spending. Also, a key benefit of having a large number of data sets available is the ability to fuse information Barbosa *et al*. [2014];
- Most commonly used formats of available data, by platform: aiming to verify whether the open definition[8] requirement — machine-readable — has been met, and also whether it was in open formats[9]. Machine-readable states that the work must be provided in a format that is easily processable by a computer and where the individual elements of the work can be easily accessed and modified.

5.1 Quantitative analysis

Through **Groups/Themes by platform**, it is possible to find sets of data classified by themes; for example, the group ‘Government and Politics’ of the Brazilian open data portal brings data related to legislative censuses, data on the organizational structure of the federal executive branch, etc. The diversity of groups, as long as it is well structured, leverages the correct use of data, in addition to helping to map which groups are pending registration and encouraging those responsible for them to register them on the portal.

According to the **Figure 3**, of the analyzed platforms, respectively, the one with the most groups is the state of São Paulo (SP) with 39 groups, followed by Espirito Santo (ES) with 26, Rio Grande do Sul (RS) with 25, Santa Catarina (SC) with 25, Bahia (BA) with 21, Mato Grosso do Sul (MS) with 12, Distrito Federal (DF) with 11, Alagoas (AL) with 10, Minas Gerais (MG) with 4, Pernambuco (PE) with 3 and Goiás (GO) with 0 groups counted.

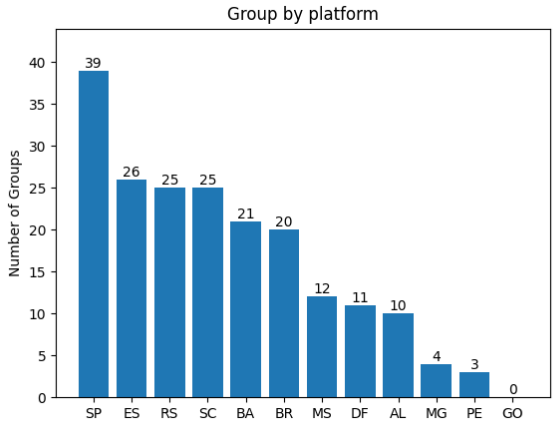

Figure 3. Quantity of groups per State and Federal District. Source: Author

Excluding the largest and smallest values (smallest value

[7] http://opendatahandbook.org/guide/en/how-to-open-up-data/#make-data-available-technical-openness

[8] https://opendefinition.org/od/2.1/en/

[9] https://opendatahandbook.org/guide/en/appendices/file-formats/

greater than zero), the platforms have an approximate average of 17,7 registered groups, and the biggest group is almost six times bigger than the smallest group.

5.1.1 Number of datasets per platform

Datasets are the available data by each platform/group. For example, the Brazilian federal government has datasets on the Annual Declaration of Use of Water Resources—DAURH, the number of Basic Health Units under construction, and Federal Government Public Procurement.

As important as a wide variety of groups (or themes) in a data portal is, the number of datasets per group is also important, as it provides information on the chosen theme. This enables the use of group data for various purposes, ranging from transparency to the implementation of new applications for society.

According to the **Figure 4**, of the analyzed platforms, the ones with the largest number of datasets, respectively is Sao Paulo (484), Alagoas (332), Santa Catarina (302), Rio Grande do Sul (302), Distrito Federal (161), Espirito Santo (95), Mato Grosso do Sul (44), Goias (33), Pernambuco (27), Minas Gerais (26), Bahia (11). Excluding the largest and smallest value (smallest value greater than zero), on average, platforms have 1322 datasets. And the largest dataset is almost 12 times larger than the smallest dataset. It is also possible to verify that even though Espirito Santo has more variety of groups (**Figure 3**), in contrast, Alagoas has more datasets available.

In Barbosa *et al*. [2014]'s study about the open urban data in North America, the correlation between the number of datasets and the population of each city was verified through Spearman's rank correlation coefficient (p score). The results showed that the p coefficient was 0.81, indicating a strong correlation between the number of datasets and their population. However, in this work, the same calculation was made, and the coefficient was -0.11, indicating that there was no correlation between these two variables. But, when making the correlation with *per capita income*, the correlation was 0.89, indicating a strong correlation between the city's *per capita income* and the number of available datasets.

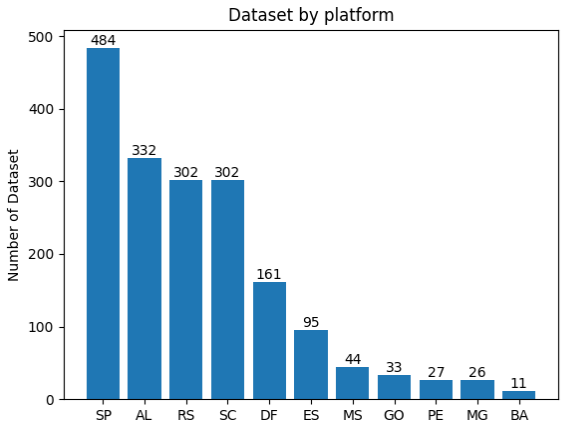

Figure 4. Quantity of datasets per State and Federal District. Source: Author

5.1.2 Most commonly used formats of available data

According to The Global Open Data Index - which is an annual effort to measure the state of open government data around the world - one of the metrics used in evaluating data sets is whether they are machine-readable, considering the types accepted, such as the following: Microsoft Excel file format, a spreadsheet file format abbreviated XLS; comma-separated values, the file is a text file that has a specific format which allows data to be saved in a table structured format abbreviated as CSV; JavaScript Object Notation, more commonly known by the acronym JSON, that is an open data interchange format that is both human and machine-readable; e XML extensible markup language, that is markup language, that describes the text in a digital document.

In the quantitative analysis, the total number of readable datasets in the general scope was verified by crossing all the portals. The result in **Figure 5** shows that 70.45% of the data is in format CSV, 29.08% of the data is in format JSON, and finally, 0.46% of the data is in format XLS.

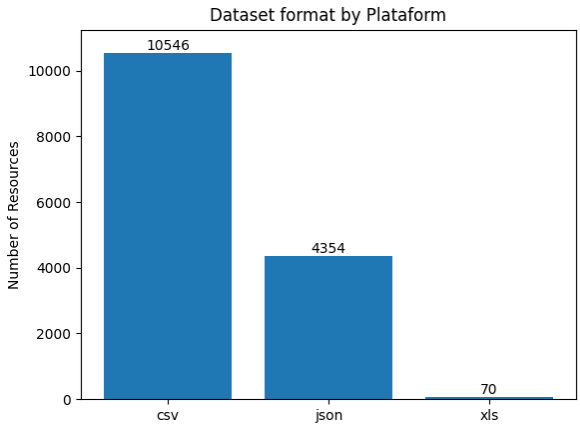

Figure 5. Most commonly used formats of available data. Source: Author

5.2 Qualitative analysis

Each subsection in the next sections will represent the OKFN's analysis GOD [2023] standardization, specified in Session 5, for the dimensions that contain enough available datasets for reproducibility and comparability purposes. The government agencies that contain the data of the analyzed dimensions can be seen on **Table 4** and will have their analysis shown in the following sections.

The dimensions (which can also be understood as categories, according to the Open Knowledge Foundation's Global Open Data Index project) on **Table 4** reflect important data that is relevant to civil society in general. In this scenario, Brazil has institutions and agencies that are, at the national/federal level, responsible for providing such information to Brazilians. Therefore, the **Table 4** lists the official government agencies that have the domain and obligation to provide correct and official information of that scope. Therefore, officially, only the agencies listed by category can provide such information to the population, so it would not be up to other agencies to publish data not related to their own scope.

Dimension	Government Agencies
National Statistics	IBGE/IPEA
National Map	IBGE
Legislation	FEDERAL SENATE
Election Results	TSE
Government Budget and Spending	CGU/TCU/BNDES
Pollutant Emissions	MMA/IBAMA
Company Register	DREI
Location datasets	
Government procurement tenders (past and present)	ME
Water Quality	ANA
Weather forecast	INMET
Land Ownership	INCRA

Table 4. Government Agencies by Dimension

Due to the characteristic of the study - analysis in open data portals only - the items/questions: *'Is data in digital form?'*, *'Publicly available?'*, *'Is the data available for free?'*, *'Is the data available online?'* of the session 5, have the same affirmative answer in common, which is: *all dimensions where the data exists were found in the open data portals: either in the Brazilian open data portal or in the government's open data portal*. Therefore, the common answer for dimensions, such as National Statistics, National Map, Election Results, Government Budget and Spending, Company Register, Location datasets, Government procurement tenders, Water Quality, Weather forecast, and Land Ownership, is that the data is available digitally. Also, it is publicly available (*on new Brazilian open data portal, it is necessary to register and then login*), and finally, it is available online and for free.

5.2.1 Does the data exist?

For the analysis *'Does the data exist?'*, all dimensions analyzed - except *'Location datasets'* and *'Company Register'* - have their own open data portal or expose their data on the Brazilian open data portal. That is, in 83% of the dimensions analyzed, the data exists.

For dimensions *Company Register*, the public agency is DREI (an institute part of the Ministry of Economy), which does not have its own open data portal, nor does it have data available on the Brazilian open data portal. However, DREI has its own website, with plans to open data linked to the agenda of the Ministry of Economy, which has not yet been implemented. For the dimension *Location datasets*, no open data were found in any portal or website of the public agency that could represent the dimension.

5.2.2 Is the data machine-readable?

In order to validate this category, the format of the data found was checked and summed - as long as they comply with the definition of the Global Open Data Index for machine-readable formats, that is, that are of types <'XLS', 'CSV', 'JSON', 'XML'>.

Some government organizations, such as *Federal Senate*, had few (under 100) datasets available on the Brazilian open data portal or on its own portal. Scenarios that can cause this include the organization's not having data exposed through the CKAN API or having data available in another non-standard model, making systematic analysis impractical. This will cause the result to be irrelevant to the dimension under analysis. When this happens, it will be signaled in the text.

- National Statistics and National Map: In total, 430 datasets were evaluated, 424 from IBGE and six from IPEA. Of these, 373 were CSV, which is equivalent to 86.74% of the formats available for this dimension, followed by the JSON format with 52 available resources and XML with 49;
- Legislation: Available to the Federal Senate, there were only two datasets and three resources, two of which were machine readable: One CSV and one XLS. In cases where the volume of data is insignificant, percentages will not be provided so that the reading is not biased, since in this case, the rate of the total resources readable by the machine would be high, but this happens because the total amount of available resources is low;
- Election Results: In total, 144 datasets were evaluated from TSE, and for these, 169 resources were used. Of the available resources, 95 were CSV, equivalent to 64.37% of the resources available for this dimension. Of the machine-readable formats, only CSV was available;
- Government Budget and Spending: In this dimension, the government agencies analyzed were The National Bank for Economic and Social Development (BNDES in Portuguese), the Federal Comptroller's Office (Portuguese acronym CGU), and the Federal Audit Court (Portuguese acronym TCU). For the three institutions, there were 92 datasets available, with 69.56% of CSV. This absolute quantity is 64, with 64 resources available, followed by 1 XML, 1 XLS, and 1 JSON;
  Of the three institutions, the BNDES stands out most positively, with 93 resources available in its 46 datasets. 47 of the 93 available resources are machine-readable. This means that, of the total resources made available by the BNDES, 50.54% can be used in automated studies with the aid of software;
- Pollutant Emissions: The government institutions analyzed were the Ministry of Environment and Climate Change (MMA) and the Brazilian Institute of Environment and Renewable Natural Resources (IBAMA). There were 98 datasets available, and, for these, 314 resources, 212 of which are machine-readable. In total, for this dimension, 67.51% of the available resources could be used systematically;
- Water Quality: The National Water and Basic Sanitation Agency (ANA) represents this dimension. It has its open data portal, with 15.46% of the total 1611 resources available, readable by machine. The predominant file type is CSV, and the total available datasets is 308;
- Weather forecast: for this dimension is the National Institute of Meteorology (INMET) of the Ministry of

Agriculture and Livestock. Which provides the following machine-readable data formats, with their respective amount: 'CSV': 27, 'XML': 14, 'JSON': 12. This means 55.79% of reusable resources, with a total of 95, spread across 56 datasets;
- Land Ownership: The National Institute of Colonization and Agrarian Reform (INCRA) will represent this session. However, only 1 dataset and resource was available, and it was not machine-readable.

5.2.3 Available in bulk?

To validate that the data was available in bulks, all resources available for the dimension under analysis were collected, and a GET-type request was made to the endpoint of each resource. Whenever the request returned the *HTTP 200* status, the resource was counted as accessible, and when it returned a different *HTTP* status, or the query exceeded the 15 seconds timeout limit to return a response, it was counted as not available resource.

- National Statistics and National Map: IBGE and IPEA together have 595 resources, of which 504 returned a success status, and 91 either reached the maximum response waiting limit of 15 seconds or returned a status other than 200. Of the available bulks, 84.70% were returning success status *200 OK*, which indicates that the request was successful;
- Legislation: The Federal Senate has 16 resources, and all returned success status;
- Government Budget and Spending: CGU, TCU, and BNDES have 334 resources, and accessible bulks are 333, only 1 with or with unavailable or failed bulk. Then, 96.80% were returning success status *200 OK*;
- Pollutant Emissions: MMA and IBAMA have 335 resources, 128 accessible, and 217 (all from IBAMA) with unavailable or failed bulk. So, only 38.20% were returning success status *200 OK*;
- Water Quality: ANA has 249 resources, and accessible bulks are 240. Of these nine, either with unavailable or failed bulk. So 96.38% were returning success status *200 OK*;
- Weather forecast: The INMET has 325 resources, of which 293 returned success. That is, 90.15% were returning success status *200 OK*;
- Land Ownership: INCRA has no resources available.

5.2.4 Openly licensed?

To validate this category, the list of licenses available for the organization under analysis was consulted in the metadata. Not all public agencies/institutes analyzed provided this information, so the number of available licenses may be smaller than the number of available datasets but never greater since the license is assigned to a dataset (and not to the multiple resources a dataset can contain).

- National Statistics and National Map (IBGE/IPEA): For this dimension there were 372 datasets with license type *'License not specified'*; 36 datasets licensed *Other (Public Domain)*; followed by *'Creative Commons Attribution'* *'Open Data Commons Open Database License (ODbL)'* *'Other (Open)'*, each with six datasets and finally four datasets with *'Other (Assignment)'* license. Of the total number of licenses reported for this dimension, 86.51% have a 'not specified' license;
- Legislation (Federal Senate): For the only one dataset available, the license was *'Creative Commons Attribution and Share Alike'*;
- Election Results (TSE): For the 144 datasets available, 140 or 97.22% use the license: *'Creative Commons Attribution'*, the only one available for the dimension;
- Government Budget and Spending (CGU/TCU/BNDES): The available licenses per dataset were *'Open Data Commons Open Database License (ODbL)'*: 46, *'Unspecified License'*: 18, *'Other (Open)'*: 13, *'Creative Commons Attribution'*: 9, *'Other (Public Domain)'*: 5. Once again, the BNDES institution stands out as responsible for making its total datasets available, under the *'Open Data Commons License (ODbL)'* license;
- Pollutant Emissions (MMA/IBAMA): The licenses are: *'Creative Commons Attribution'*: 36, *'Other (Public Domain)'*: 35, *'Other (Open)'*: 20, *'Open Data Commons Open Database License (ODbL)'*: 7. In total there are 98 specific licenses for the 98 available datasets;
- Water Quality (ANA): For this dimension, the metadata query did not return available licenses;
- Weather forecast (INMET): For this dimension, all 56 available datasets have the following licenses: *'Creative Commons Attribution'*: 48, *'Open Data Commons Open Database License (ODbL)'*: 6, *'Other (Open)'*: 2;
- Land Ownership (INCRA): For the only available dataset, the license is *'Creative Commons Attribution'*.

5.2.5 Is the data provided on a timely and up-to-date basis?

For this category, a search was made in the metadata of the dataset, if they had the keywords: "update frequency", "update frequency (months)", "periodicity", and "publication frequency". As this attribute is not mandatory, some governmental organizations did not place it, or if they were placed, it was not in a standardized way, so that, if there was some periodicity informed that does not meet the keywords mentioned at the beginning of the paragraph, this was not accounted for.

Understanding the importance and relevance of the information, manual searches were performed for more diversity of keywords that could somehow contain the periodicity. However, as mentioned, it was a non-mandatory attribute and, therefore, not standardized in how it was made available. Due to the volume of data, a manual search is not effective and sufficient. When the periodicity information was informed, it was then verified if the value was in the following list: ['biannual', 'semi-annual', 'daily', 'weekly', 'bi-weekly', 'monthly', 'bimonthly', 'quarterly ', 'annual', 'bi-ennial']. If yes, the value of the ['metadata modified'] attribute was reduced from the founded value in the list. For example, dataset X contained a 'monthly' update periodicity, so it was checked if the last dataset update had been done in the last month. If the value returned for the periodicity was

anything other than the one in the list above, such as <update frequency: 'on demand'>, the calculation was impossible, and the dataset could not be counted.

- National Statistics and National Map (IBGE/IPEA): IPEA did not return - in its metadata - freshness data of the datasets. As for the IBGE, of its 424 datasets, 53 had freshness information, of which only 2 were up to date according to the calculations made between the informed update window and the last modification of the dataset. That is 3.77% of up-to-date datasets, of the total, that contained freshness data;
- Legislation (Federal Senate): For the only available dataset, there was no freshness information on metadata;
- Election Results (TSE): Of its 144 datasets, 20 had information about freshness, and none - according to the informed update window - was up to date;
- Government Budget and Spending (CGU/TCU/BNDES): For TCU it did not contain new information in its datasets. For CGU, of its 43 datasets, 12 had freshness information, that is, 27.91%. Of these, none of the 12 were up to date. As for the BNDES, 58.7% of its datasets had freshness information. That is, out of a total of 46, 27 had metadata with freshness information. And of the 27 with freshness information, all were up to date;
- Pollutant Emissions (MMA/IBAMA): For MMA, 6.06% contained freshness information - 4 out of 66 datasets reported freshness metadata - of these, 25.0% or only 1 was up to date. This scenario should be looked at carefully because despite the percentages of updates being relatively considerable, the amount of datasets available in total is under 100, which is a low amount. For IBAMA, 56.25% contained freshness information - 18 out of 32 datasets reported freshness metadata - of these, 18 were up to date;
- Water Quality (ANA): Despite containing a relevant amount of datasets - if considered with other government bodies - 300, the ANA agency returned that 0 datasets are up to date, which drew attention during this study. Occasionally, in this case, due to the quality of the ANA results in other analyses, it was manually verified that, in fact, this field does not exist in the metadata of the datasets, so in this case, it was not possible to verify if the data is updated or not;
- Weather forecast (INMET): INMET has a total of datasets: 56, of which 13 or 23.21% with freshness information. Only 7.69% or 1 dataset was up to date;
- Land Ownership (INCRA): INCRA has only one available dataset and no freshness information.

6 Conclusions

The quantity analysis validated 1817 datasets from 196 groups contained in nineteen open data portals, which are subdivided among twelve open data portals of the states, two open data portals in Brazil and the Federal District, and five government agency portals with their open data portals, such as TSE, BNDES, MMA, IBAMA, and ANA.

Although Brazil has 26 states, less than half, only 11, have an open data portal with standard exposure using CKAN - widely used in countries that expose their data. However, of those who do not expose their data with standardization, even though some, such as Rio de Janeiro, have an open data portal, this demonstrates a willingness and understanding of the importance of this subject.

Although states without portals with standards demonstrate knowledge of the importance of open data, this highlights one of the main difficulties reported in the study, the lack of promotion and dissemination of the importance of having standards in data exposure so that it is possible to reuse data, or apply systematic studies, for example, through multiple data portals in different geographic locations.

The analysis of dimensions was important for a deeper dive into the data, according to important issues for the population, such as pollutant emissions and national statistics. For this category, manual searches were carried out to verify the existence of open data portals of federal government agencies nationwide responsible for making these services available to Brazilian citizens. For example, the federal bodies IBGE and IPEA represented the national statistical dimension, and the pollutant emission dimension was represented by the institutions MMA and IBAMA.

It was perceived that important federal government organizations, such as the Federal Senate, despite having a transparency portal, do not have their own open data portal (different purposes). The analyzed data was accessed via the Brazilian open data portal, and with a number of datasets less than 100, it is a low quantity compared to other organizations such as IBGE, which contains more than 400.

Regarding the category *"is the data machine readable"* of the dimensions that had a relevant volume of data (excluding Legislation and Land Ownership), all had at least 50% of their data readable by machines, that is, capable of reuse. The Water Quality dimension was the exception with 15%. The category *"Available in bulk?"* of the dimensions that had data (excluding Land Ownership), all had at least 80% of their resources successfully accessed. Despite the high volume, 80%, of resources available in bulks, the value needs to be looked at carefully, as having 80% of the data available does not mean a relevant volume of resources, it just means that, from those made available, there was a high possibility of success in the access.

The last two categories evaluated were those with the worst indicators for all dimensions, representing the difficulties read in related materials, which is: unclear licensing, as, for the most part, the analyzed dimensions have licenses of the type: *'License not specified'*, *'Unspecified License'* and *'Other (Open)'*, when specified. Regarding the resources being up-to-date, in all dimensions, less than 10% of the resources proved to be up-to-date or did not have enough metadata for correct freshness analysis.

6.1 Future Work

One of the biggest concerns in our society is privacy, which is identified as a fundamental political, regulatory, and legislative challenge of the 21st century, with the scope of privacy being expanded to cover different aspects, including control

over information, human dignity, intimacy, and social relationships. The unauthorized manipulation, collection, and use of personal data raises important ethical and privacy issues Sokolovska and Kocarev [2018].

Personal data protection laws such as LGPD[10] in Brazil and GDPR[11] in Europe are in force, in order to control the availability and access of personal data on the internet, seeking to preserve people's right to privacy. Given the nature of this study, open data, the evolution of this work must contain the verification of compliance with Brazil's data protection law, LGPD, on open data portals, considering the question: *"Is there personal data exposed on portals of open data in Brazil?"* to notify public organizations if exposed personal data is found, which is a clear violation of human rights[12].

Declarations

Acknowledgements

This study was financed partly by the Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES) - Finance Code 001 and partly by the National Council for Scientific and Technological Development - Brazil (CNPq). Also, partial support was provided via the research project S2C2, ref. 2904/20 and CEREIA Project (# 2020/09706-7) São Paulo Research Foundation (FAPESP), FAPESP–MCTIC-CGI.BR in partnership with Hapvida NotreDame Intermedica group. FAPERGS Project Smart-Sent No. (17/1195-3). FAPERGS Project GREEN-CLOUD (16/2551- 0000 488-9)

Funding

This study was financed partly by the Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES) - Finance Code 001 and partly by the National Council for Scientific and Technological Development - Brazil (CNPq).

Authors' Contributions

Do Carmo - wrote the introduction, model, results, and conclusion sections, drew the figures, and computed the experiments; Dos Anjos - critical review of the methodology, data analysis model, and evaluation strategies; Geyer - proofread and supervised.

Competing interests

The authors declare they do not have competing interests.

Availability of data and materials

Available upon request to the authors

[10] https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm

[11] https://gdpr-info.eu/

[12] https://www.hrw.org/news/2018/06/06/eu-general-data-protection-regulation

References

(2023). Godi methodology. Available at: `http://index.okfn.org/methodology/`. Accessed on: 2023-10-10.

Adje, K. D. C., Letaifa, A. B., Haddad, M., and Habachi, O. (2023). Smart city based on open data: A survey. *IEEE Access*, 11:56726–56748. DOI: 10.1109/ACCESS.2023.3283436.

AnacondaOrg (2023). Package categories. Available at: `https://anaconda.cloud/package-categories`. Accessed on: 2023-10-10.

Arbex, A. M. G. (2020). Como os dados abertos podem revolucionar as cidades. Available at: `https://forbes.com.br/colunas/2020/01/como_os_dados_abertos_podem_revolucionar_as_cidades/`.

Barbosa, L., Pham, K., Silva, C., Vieira, M. R., and Freire, J. (2014). Structured open urban data: Understanding the landscape. *Big Data*, 2(3):144–154. DOI: 10.1089/big.2014.0020.

Chen, L. and Jakubowicz, J. (2015). Inferring bike trip patterns from bike sharing system open data. In *2015 IEEE International Conference on Big Data (Big Data)*, pages 2898–2900. DOI: 10.1109/BigData.2015.7364115.

CKAN (2023). Available at: `https://docs.ckan.org/en/2.9/api/`. Accessed on: 2023-10-10.

DESA, U. (2018). 68 Available at: `https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html`.

Domingue, J. *et al*. (2011). *The Future Internet*, volume 1. Springer. BOOK.

Federal, G. (2023a). "infraestrutura nacional de dados abertos". Available at: `https://www.gov.br/governodigital/pt-br/dados-abertos/infraestrutura-nacional-de-dados-abertos`. Accessed on: 2023-09-12.

Federal, G. (2023b). "política de dados abertos". Available at: `https://dados.gov.br/dados/conteudo/politica-de-dados-abertos`. Accessed on: 2023-09-12.

Foundation, O. K. (2023). What is open? `https://okfn.org/opendata/`.

Gharawi, M. A., Al Hamed, K. M., and Alneami, H. H. (2019). Compliance with open data principles: A longitudinal content analysis of the saudi's national open data platform in 2016 and 2018. In *2019 IEEE 2nd International Conference on Information and Computer Technologies (ICICT)*, pages 80–87. DOI: 10.1109/INFOCT.2019.8711298.

Guo, G., Khalil, J. M., Yan, D., and Sisiopiku, V. (2019). Realistic transport simulation: Tackling the small data challenge with open data. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 4512–4519. DOI: 10.1109/BigData47090.2019.9006457.

Index, O. D. (2018). Open data index. Available at: `https://ok.org.br/projetos/open-data-index/`.

Janssen, M., Charalabidis, Y., and Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and

open government. *Information Systems Management*, 29(4):258–268. DOI: 10.1080/10580530.2012.716740.

JupyterOrg (2023). Available at: `https://docs.jupyter.org/en/latest/install/notebook-classic.html`. Accessed on: 2023-10-10.

JupyterOrg Community (2023). Content community. Available at: `https://docs.jupyter.org/`. Accessed on: 2023-10-10.

Kirstein, F. and Bohlen, V. (2022). *IDS as a Foundation for Open Data Ecosystems*, pages 225–240. Springer International Publishing, Cham. DOI: 10.1007/978-3-030-93975-5_14.

Machado, J. S., Farah, J. C., Gillet, D., and Rodríguez-Triana, M. J. (2019). Towards open data in digital education platforms. In *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)*, volume 2161-377X, pages 209–211. DOI: 10.1109/ICALT.2019.00048.

MatplotlibOrg (2023). Visualization with python. Available at: `https://matplotlib.org/`. Accessed on: 2023-10-10.

McDermott, P. (2010). Building open government. *Government Information Quarterly*, 27(4):401–413. Special Issue: Open/Transparent Government. DOI: 10.1016/j.giq.2010.07.002.

Nogueras-Iso, J., Lacasta, J., Ureña-Cámara, M. A., and Ariza-López, F. J. (2021). Quality of metadata in open data portals. *IEEE Access*, 9:60364–60382. DOI: 10.1109/ACCESS.2021.3073455.

PandasOrg (2023). Python data analysis library. Available at: `https://pandas.pydata.org/`. Accessed on: 2023-10-10.

Pareja-Lora, A., Blume, M., Lust, B. C., Chiarcos, C., Chiarcos, C., Pareja-Lora, A., Langendoen, D. T., Ide, N., Moran, S., Warburton, K., Wright, S. E., Trippel, T., Zinn, C., Simons, G., Bird, S., Ratner, N. B., MacWhinney, B., Blume, M., Flynn, S., Foley, C., Caldwell, T., Reidy, J., Masci, J., Lust, B. C., Barrière, I., Dye, C., Kang, C., and Rieger, O. (2019). Development of linguistic linked open data resources for collaborative data-intensive research in the language sciences: An introduction. In *Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences*, pages ix–xxi. DOI: 10.7551/mitpress/10990.001.0001.

PythonOrg (2023). Content community. Available at: `https://www.python.org/community/`. Accessed on: 2023-10-10.

Sokolovska, A. and Kocarev, L. (2018). Integrating technical and legal concepts of privacy. *IEEE Access*, 6:26543–26557. DOI: 10.1109/ACCESS.2018.2836184.

Zhang, C. and Yue, P. (2016). Spatial grid based open government data mining. In *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 192–193. DOI: 10.1109/IGARSS.2016.7729041.